Grid y Computación
de Altas Prestaciones
GRyCAP

RETOS COMPUTACIONALES EN LA MEDICINA PERSONALIZADA

Ignacio Blanquer
Vicente Hernández

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

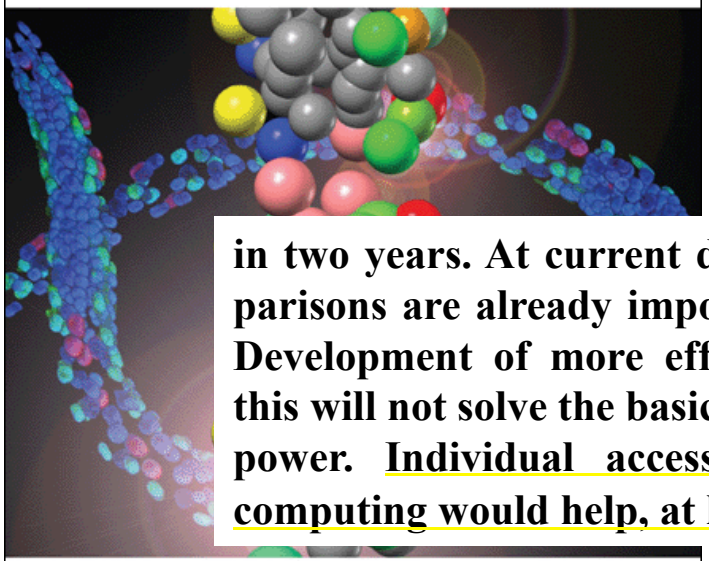
i3M | Instituto de Instrumentación para Imagen Molecular

GENÓMICA Y LAS NUEVAS TECNOLOGÍAS

- Comunidad Acostumbrada a Trabajar de Forma Colaborativa en Internet
  - GeneBank, PDB, SWISSPROT, KEGG, Portal de NCBI.
- Usuarios de Muchas Herramientas Comunes
  - BLAST, ClustalW, ePCR, SAGE, EMBOSS , Phylip, …
- Una Comunidad Abierta y Partidaria del Código Abierto y el Acceso Abierto a los Datos.
- Consumidores de ~10% de los Recursos Computacionales de las Infraestructuras Europeas de Investigación.

ESTA PERCEPCIÓN NO ES SÓLO DESDE LAS TIC

September 2009 | volume 6 | number 9
nature methods
www.nature.com/naturemethods
Techniques for life scientists and chemists



- A digital atlas of the worm
- Tools for metagenomics
- A reporter line for erythroid differentiation
- BreakDancer
- Addressing the crystallography phase problem

Metagenomics versus Moore's law

Metagenomics sprang from advances in sequencing technology, and continued improvements are providing data in quantities unimaginable a few years ago. But without concerted efforts, the amount of data will quickly outpace the ability of scientists to analyze it.

As Craig Venter sails the oceans collecting seawater samples to profile microbial communities by high-throughput sequence analysis, microbiologists around the world are busy collecting their own samples. The diversity of locations—from Antarctic lakes to human armpits—highlights the reality that microscopic organisms represent a significant fraction of the Earth's ecosystem.

Any population this large is certain to have profound influences on its environment. Yet our knowledge of

40 megabases. Today there are over 4,000 sequenced metagenomes, and their size and number are increasing. Each new pyrosequenced metagenome is 200–500 megabases, and those generated on Illumina platforms are 20–50 gigabases. To analyze these metagenomes using established pipelines would take tens of years on a single processor and weeks to months on machines with up to 1,000 processors. The rate of increase in sequence generation is far outpacing Moore's law, and the cost of analyzing the largest datasets

in two years. At current database sizes all-versus-all comparisons are already impossible without a supercomputer. Development of more efficient algorithms will help, but this will not solve the basic problem of too little comput-ing power. Individual access to supercomputers or cloud computing would help, at least temporarily.

SOLiD platforms. Most metagenomics analysis pipelines are designed for Sanger sequencing data, so the short read lengths and error profiles of the new methods present challenges for data analysis and interpretation.

Reports on pages 639 and 673 and an accompanying News and Views on page 636 illustrate some of the dangers and challenges involved and describe new algorithms to deal with them. More work is needed to assess the new technologies and develop optimized analysis pipelines, and these efforts are well underway.

But even as these problems are being solved, a larger problem has taken the community off-guard: the exponentially increasing amount of sequence data. Just over three years ago, the first two second-generation sequencing platform–based shotgun metagenomes were reported—each less than

decrease computational demands by improving data sharing through standards and centralized coordination and by aggregating computationally intensive operations.

This summer, after discussions at the International Conference on Systems for Intelligent Molecular Biology, community members formed the M5 (metagenomics, metadata, metaanalysis, multiscale-models and metainfrastructure) Consortium under the roof of the Genomics Standards Consortium to devise a solution to the coming gridlock. Their proposed 'M5 Platform'—to be announced later this year—deserves the support of the community, funding agencies and those who hold the keys to the high-performance computing centers. Unless major efforts are taken immediately, researchers will find they have a wealth of data but no way to interpret it.

INFRAESTRUCTURAS COMPUTACIONALES

Google App Engine

amazon web services™

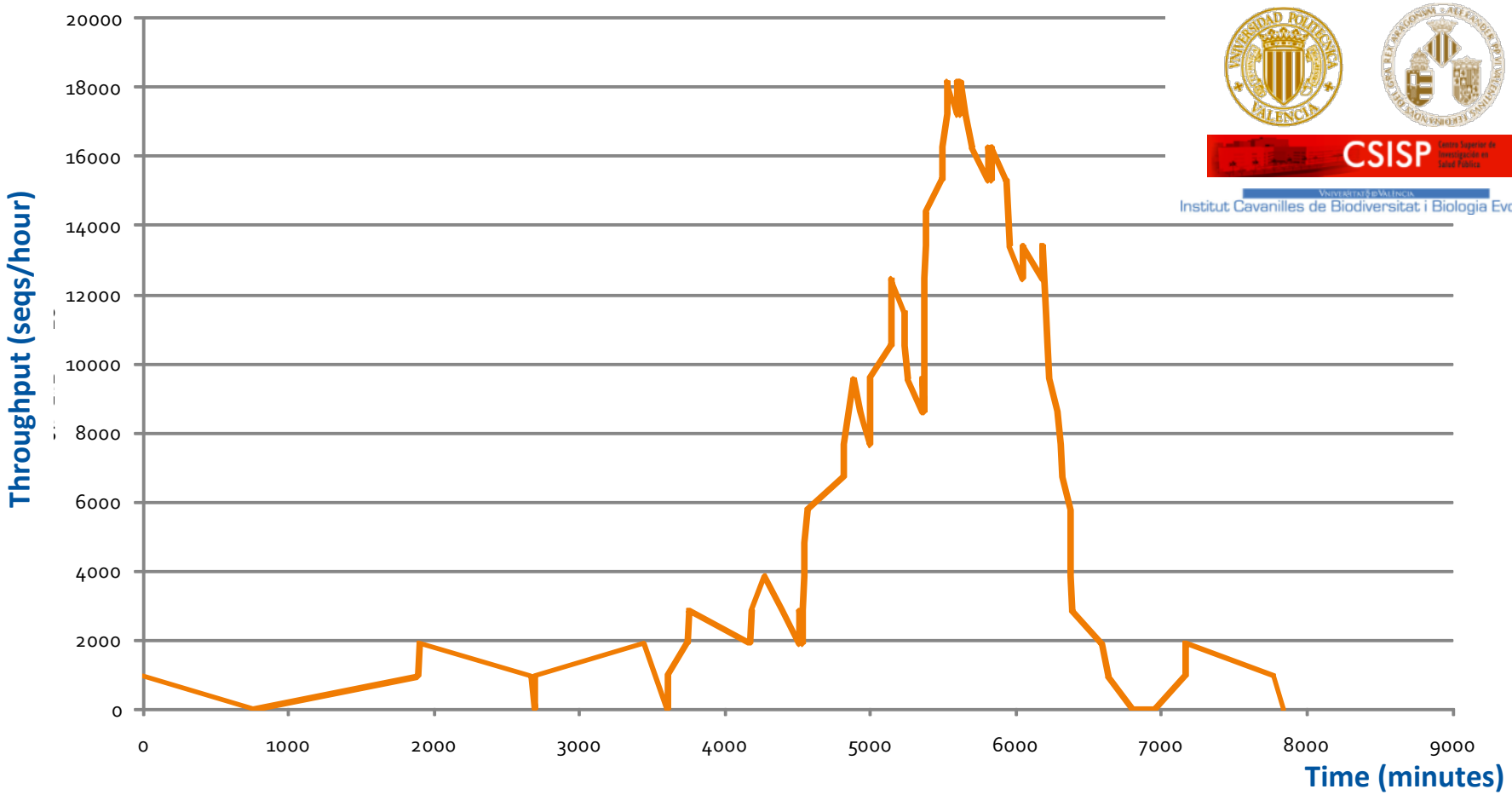
Windows Azure™

EXPERIMENTOS A GRAN ESCALA

Bioinformatics Use Case

GRyCAP
Grid y Computación de Altas Prestaciones
www.grycap.upv.es

- 600 Computadores Usados Simultáneamente en EGI.



CSISP
Institut Cavanilles de Biodiversitat i Biologia Evolutiva

DIAGEN

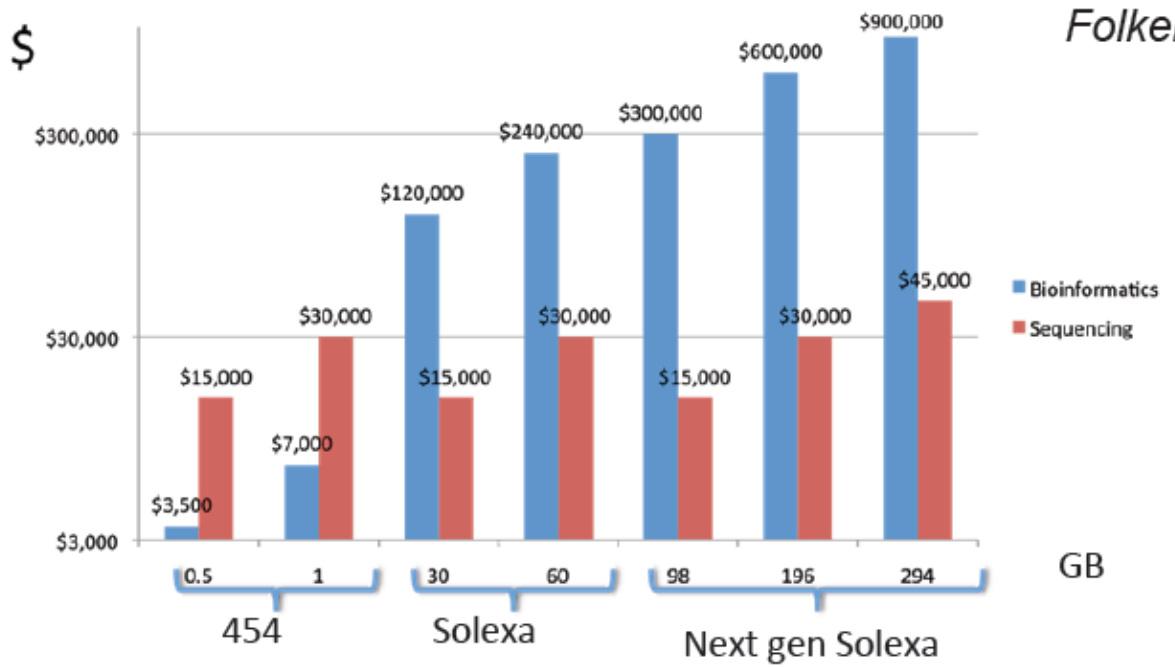
- A Pesar del Crecimiento que las Bases de Datos de Información Genética, Genómica y Proteómica no hay Cambios en la Estructura Original.
- Por ello, Tanto la Eficiencia Como la Coherencia de los Contenidos se Debilitan a Medida que su Tamaño Aumenta Exponencialmente.
- Una Aplicación Bioinformática Moderna Exige un Acoplamiento Fino entre Diseño Correcto de Datos y Mecanismos Potentes de Búsqueda.
  - Estos Contenidos Deben de Explotarse con Herramientas de Búsqueda con la Potencia Computacional Necesaria.

PROS
Centro de Investigación en Métodos de Producción de Software

Grid y Computación de Altas Prestaciones
GRyCAP

PERO LAS COSAS SE COMPLICAN

the "simple" problem: sequencing outpaces moore's law

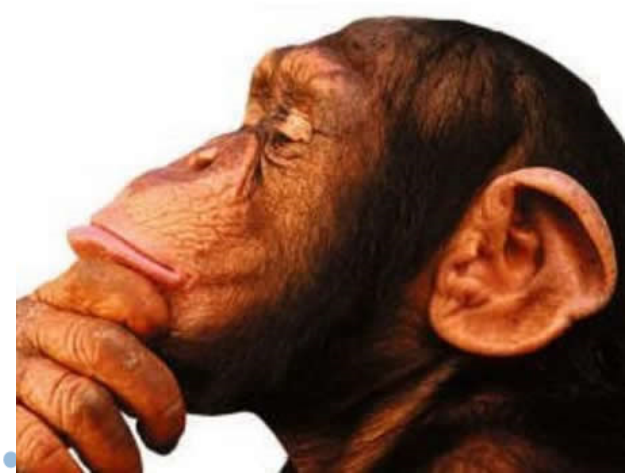


- Values are for BLAST searches on Amazon EC2 (from *Wilkening et al, IEEE Cluster09*)
- Acknowledged: BLAST is not the ideal tool for a lot of this, but…

O INCLUSO PEOR

CONCLUSIONES

- La Genómica puede Convertirse en la Actividad con Mayor Demanda Computacional y de Almacenamiento en la Industria y la Investigación.
- Las Técnicas de Cloud Computing se ven Como una Solución, al Menos a Corto Plazo.
- Las Infraestructuras Públicas de Investigación y los Proveedores Comerciales Jugarán un Papel Fundamental en la Medicina Personalizada.



CONTACTO

GRyCAP
Grid y Computación de Altas Prestaciones
www.grycap.upv.es

Ignacio Blanquer
Vicente Hernández
Universidad Politécnica de Valencia
Camino de Vera s/n,
46022 Valencia
Tel: +34-963879743
Fax. +34-963877274
E-mail: iblanque@dsic.upv.es
vhernand@dsic.upv.es